Addressing Technical Replicate Variance in Omics Data Analysis

Enrico Glaab and Reinhard Schneider
Contact: enrico.glaab@uni.lu

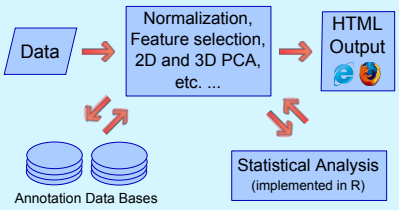
www.repexplore.tk

1 Introduction

Omics datasets often contain technical replicates, included to account for technical noise in the measurement process. Summarizing these replicates using robust averages may help to reduce the influence of noise on downstream data analysis, but the information on the variance across replicate measurements is lost for subsequent analyses. We present **RepExplore**, a web-service to **exploit the information captured in the technical replicate variance** to provide more robust differential abundance statistics and principal component analyses for omics datasets. A **fully automated data processing pipeline** and **interactive ranking tables** and **2D and 3D visualizations** further facilitate the interpretation of complex experimental data.

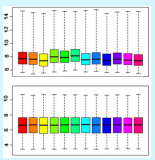
2 Workflow

Analyzing omics data with RepExplore requires only the upload of a tab-delimited dataset containing both technical and biological replicates for different conditions of interest. Alternatively, all functions can be tested with previously published example data. The input is processed automatically, including optional normalizations, and the results are combined into a single web-based report for interactive exploration.

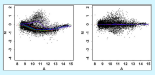


3 Data Normalization

On RepExplore, the user can either provide fully pre-processed data as input or let the software apply different **automated and parameter-free normalization procedures**. For example, RepExplore can automatically adjust the scaling of samples to facilitate the comparison of data from different batches. Common and unwanted dependencies between the signal variance and average signal intensity in experimental data from high-throughput measurement platforms can also be removed without manual inspection.



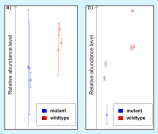
Scaling: before & after



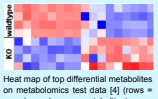
Stabilizing: before & after

4 Differential Analysis

Significant differences in the measured abundance of proteins, metabolites or mRNA transcripts between target and reference conditions are quantified robustly by **accounting for the variance in technical replicates** using the Probability of Positive Log-Ratio (PPLR) statistic [1]. For comparison, results on the mean-summarized replicates are generated additionally by applying the widely used empirical Bayes moderated t-statistic [2]. A **sortable ranking table**, **whisker plots** and **interactive heat map visualizations** enable a detailed exploration of differential abundance patterns in complex biological datasets.



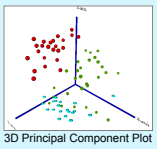
Whisker plots of top-differential metabolites on Arabidopsis thaliana test dataset [3]: a) High overlap between technical replicate variance in classical analysis; b) No overlap and small replicate variance in PPLR analysis



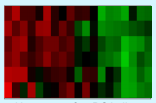
Heat map of top differential metabolites on metabolomics test data [4] (rows = samples, columns = metabolites)

5 Denoised Visualization

Technical noise in high-throughput experimental data does not only affect derived statistics but also popular dimension reduction approaches for data visualization like Principal Component Analysis (PCA). By using a **generalization of Probabilistic PCA** [4] we can account for measurement uncertainty captured via technical replicates and obtain improved PCA visualizations and tighter sample clusters. RepExplore provides both **2D PCA plots and interactive 3D PCA visualizations** [5] which exploit the information on measurement variance for each biomolecule.



3D Principal Component Plot



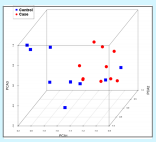
Heat map after PCA dimension reduction (columns = samples, rows = genes)

6 Web-service Automation

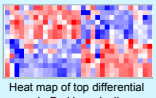
To facilitate and speed up the analysis of large numbers of datasets, the software can be accessed via an **exposed programmatic web-service API**, enabling users to submit analyses from a wide range of programming or scripting languages. Example scripts for an efficient and automated analysis of multiple omics datasets are provided on the RepExplore web-page.

7 Biological Results

We have tested RepExplore on proteomics and metabolomics data from published disease-related case/control studies and wild-type/knockout studies. Compared to the standard approach of applying a differential abundance statistic to mean-summarized technical replicates the **value ranges of identified top differential biomolecules display smaller or no overlap across the sample groups** and the overall replicate variance is significantly smaller. The Probabilistic PCA provides **improved low-dimensional data visualizations with a tighter clustering of samples**.



Probabilistic PCA of Parkinson's disease data



Heat map of top differential genes in Parkinson's disease

8 Conclusion

RepExplore is a free web-service for transcriptomics, proteomics and metabolomics data analysis providing:

- **improved robustness by addressing technical replicate variance**
- **automated data processing on an easy-to-use web-interface**
- **interactive visualizations and ranking tables to explore the results**
- **fast analysis of multiple datasets via an exposed web-service API**

References

[1] Liu, X. et al. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, 22 (17), 2107–2113
[2] Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3(1), 3
[3] Anderson, J. C. et al. (2014) Decreased abundance of type III secretion system-inducing signals in Arabidopsis mkp1 enhances resistance against pseudomonas syringae. *Proc. Natl. Acad. Sci. U. S. A.*, 111 (18), 6846–6851
[4] Böttcher, C., et al. (2009) The multifunctional enzyme CYP71B15 (Phytoalexin Deficient 3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana. *Plant Cell*, 21 (6), 1830–1845
[5] Sanguinetti, G. et al. (2005) Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics*, 21 (19), 3748–3754
[6] Glaab, E., et al. (2010) vrmlgen: An R package for 3D data visualization on the web. *J. Stat. Soft.*, 36 (8), 1–18